Non-Laboratory-Based Risk Factors for Automated Heart Disease Detection

H. Mai[1], T. T. Pham*[1,2], D. N. Nguyen[2], E. Dutkiewicz[2]

[1] Garvan Medical Institute, NSW, Australia.
[2] Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia.

Abstract— **Developing a heart disease detection model using simple non-laboratory risk factors plays an important role in preventive care, especially for high risk subjects. The model allows physicians/epidemiologists to effectively diagnose a person as having heart disease. In this work, we aim to develop a non-invasive risk prediction model for automated heart disease detection that involves age, gender, rest blood pressure, maximum heart rate, and rest electrocardiography. We examine four public datasets from 1071 participants who were referred for a special X-ray of the heart's arteries (i.e., to see if they are narrowed or blocked). The subjects also undertook a physical examination and three non-invasive tests. To estimate the heart disease status, we apply a generalized linear model with regularization paths via coordinate descent. Even without laboratory-based data (e.g., serum cholesterol, fasting blood sugar), we observed a prediction accuracy as high as 72%, compared with 76% of other comprehensive models. This observation suggests that few non-invasive factors utilizing recent advances in data analytics can replace the current practices of heart disease risk assessment.**

Index Terms— **Heart disease, RBP, RestECG**

I. Introduction

Heart diseases and conditions involving the heart and blood vessels (i.e., cardiovascular disease, CVD) are the most common ones leading to death. It has been reported that 17.7 million deaths from CVDs in 2015 accounted for approximately 31% of all deaths worldwide [1]. About half of that were due to coronary heart disease [1].

Early detection of heart disease and appropriate management are critical to people with heart disease [1]. One main reason is that most CVDs can be prevented by adjusting lifestyle (e.g., smoking, significant alcohol consumption, unhealthy diet, and physical inactivity). However, current practices to detect heart disease in population still have challenges in terms of expenditure and facility availablity.

CVDs have been often diagnosed through several laboratory experiment results, e.g., blood tests, chest X-ray, or clinical assessments of electrocardiogram (ECG), echocardiogram. Recently, in a follow-up study cohort of 6186 people (over 21 years) [2], a non-laboratory-based model was shown comparable to a laboratory-based when predicting first-time fatal heart disease events. In the non-laboratory-based model, cholesterol was replaced by body-mass index. Another similar observation found in the LIPID study [3], non-laboratory-based risk factors were significantly associated with the risk of developing a recurrent heart disease event [3]. However, these approaches still require "complicated" information such as diabetes and current hypertension treatment data as in the earlier work [2] or angina grade and myocardial infarction history as in the work [3]. We hypothesize that fewer non-laboratory-based risk factors such as age, gender, rest blood pressure, maximum heart rate, rest ECG ST-segment abnormality may also have comparable performance, especially in automated heart disease detection.

Simplifying risk assessment tool is a critical step for prevention strategies. Among cardiovascular deaths, 80% occur in developing countries where laboratory-based risk tools are costly and not practical because lack of facilities. Meanwhile, non-laboratory-based information is easier and less costly to collect. Moreover, simple analysis of data like short ECG measurements can be automatically implemented with nearly similar accuracy to manual clinical assessment. On the other hand, using a very large number of subjects, accurate probability models for CVD detection could be derived and applied universally. Early efforts [3]–[5] introduced probability algorithms built from several thousand patients. A recent study [6] reported clinical data alone are insufficient to predict the disease. Meanwhile, Hemingway et al., [7] proposed to use linked electronic medical records to increase the accuracy of coronary artery disease diagnostics. Our study, in line with the application of data science to health care, aim to propose a model for automated CVD detection using only five non-laboratory factors: age, gender, rest blood pressure (BP), maximum heart rate (MaxHR) achieved, rest ECG analysis.

The main contributions of this work are:

- We utilize a maximum collection size of a large well-known dataset for heart disease from multi-nation sites [4]. In the literature, due to clinical factors suffered from missing data across different hospitals, most related works on this dataset could only use one of the four databases of the dataset (i.e., Cleveland database [4]).
- We demonstrate that our proposed model includes fewer non-laboratory-based risk factors but performs comparably to models involving laboratory-based data inputs.

II. Methods

A. Data Set

Our study used four public databases contributed by an international collaboration concerning heart disease diagnosis [4]. There are 303 instances from Cleveland Clinic Foundation, 294 from Hungarian Institute of Cardiology, 123 from

*: correspondence thuy.pham@uts.edu.au. Faculty of Engineering and IT, University of Technology Sydney, NSW, Australia

V.A. Medical Center (California, USA) and 200 instances from University Hospital (Switzerland) [4]. In general, 1071 participants who were referred for coronary angiography undertook a physical examination and three non-invasive tests. All participants had no history of myocardial infarction, valvular or cardiomyopathy disease. The class distribution of the predicted attribute among four hospitals is presented in Table I. In this work, towards an automated heart disease presence detection, we categorised them into two groups *Negative* (i.e., Class 0 in the dataset) and *Positive* (i.e., any of Class 1 to 4 in the dataset).

There are 13 risk factors provided in the dataset including laboratory-based and non-laboratory-based risk factors. Variables obtained from clinical test include chest-pain, serum cholesterol, resting blood pressure (in mm Hg) and fasting blood sugar (FBS). Whereas, non-invasive tests provide information about maximum heart rate (MHR), the slope of the peak exercise. Table II depicts the baseline characteristics of all 13 factors for 411 people who were diagnosed negative with CVDs and 509 patients who were positive. Patients with the disease appeared to be older, more likely to be men, and had a higher rest blood pressure, lower cholesterol than those in control group. People with positive CVDs also tend to have fasting blood sugar greater than 120 mg/dl. The index of all 13 selected variables was statistically different in two groups.

B. Generalized Linear Models with Regularization Paths via Coordinate Descent

To classify a subject has heart disease or not we use a generalized linear model (GLM) with convex penalties [8]. This is a binary classifier based on logistic regression (i.e., a quadratic approximation to the log-likelihood). There are three common penalties to generalize the model: $l1$ (the Lasso), $l2$ (ridge), and mixtures of the two (the elastic net) [8]. While the former does the shrinkage and variable selection at the same time, the latter may not select any subset of variables (i.e., may include all or none of them). These two approaches have different assumptions on the relationships between input and output data. For example, the ridge refers to a normal distribution for the coefficients of the linear transformation while the Lasso refers to the Laplace distribution. Thus, in this work, we compare both two models with different penalties: *Ridge* and *Lasso*. Let λ be the regularization parameter (i.e., control the weight of penalty).

C. Performance Metrics

The accuracy of proposed model in heart disease detection is evaluated as follows. Subjects who were labeled the same as annotation of *positive* are True Positives (TP). Subjects who were labeled as Positive but did not agree with the ground truth are False Positives (FP). Subjects who were labeled as Negative by the proposed method but were annotated as Positive are False Negatives (FN). The subject that was labeled as Negative by both are True Negative (TN). The sensitivity was calculated as $\frac{TP}{TP+FN}$ and the specificity was calculated as $\frac{TN}{TN+FP}$.

III. Results

1) Fitting GLM models: Fig. 1 visualizes the coefficients of fitted models using GLM approaches. In the figure, each variable is illustrated by a curve line against the l1-norm when varying λ values. The top axis represents the number of non-zero coefficients at the current λ (or can be referred to as degrees of freedom (df) for the Lasso).

2) Model Selection with Cross-validation: We implemented a ten-fold cross-validation using the misclassification error criterion. The grid of λ extends to a range of 100 values. Fig. 2 illustrates the cross-validation curve with the standard deviations (i.e., error bars). In the figure, two vertical dotted lines present two selected λ: λ_{min} (i.e., λ that gives the minimum error) and λ_{SE} (i.e., for one standard error range of the minimum error). These (λ_{min}, λ_{SE}) values, found through the cross-validation, are (0.03611, 0.1005) and (0.13598, 0.8741) for the Lasso-based and the Ridge regularization, respectively.

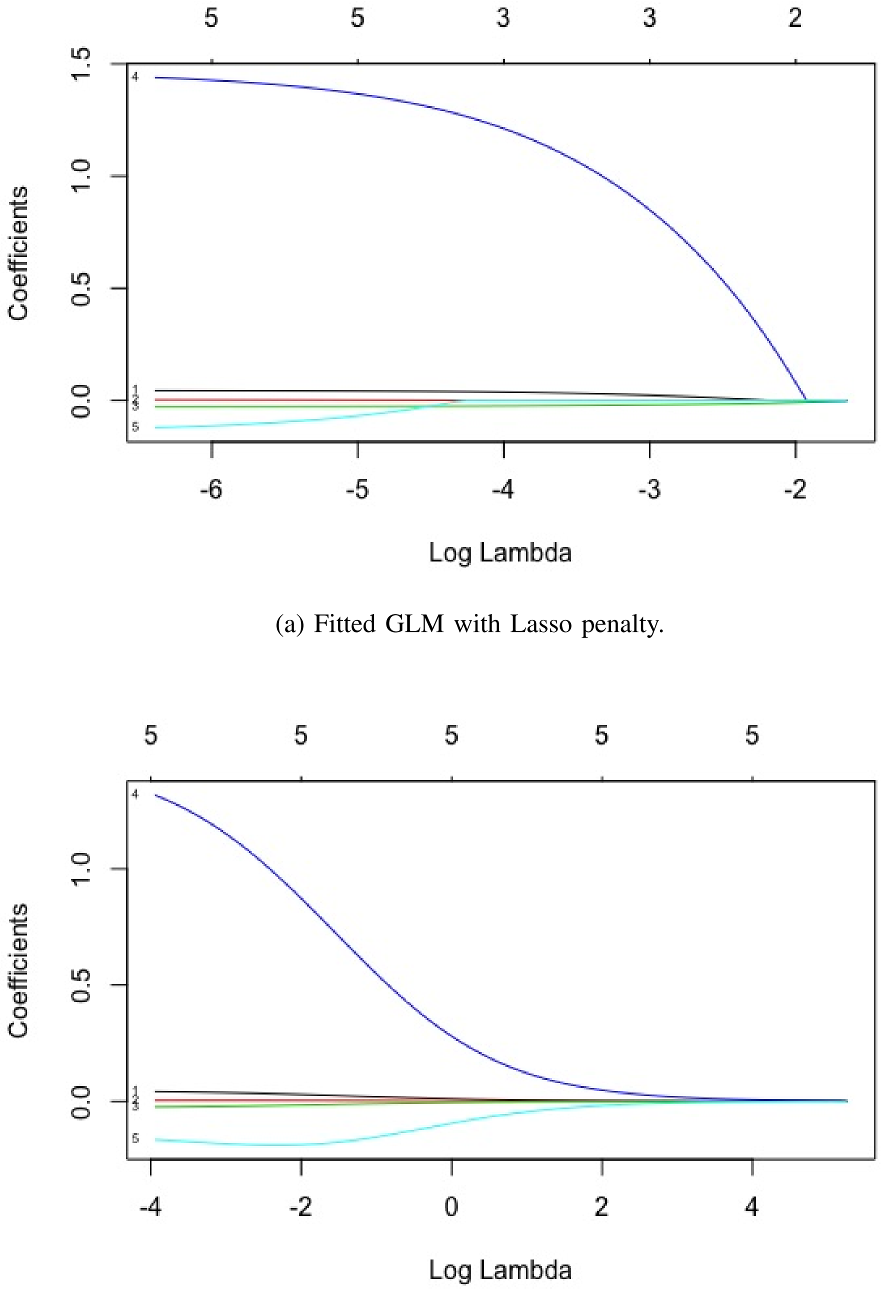
3) Selected GLM models: Table III lists the corresponding coefficients of fitted models at the best cross-validation parameter λ_{min} for five non-laboratory-based risk factors. We noticed that in the model using the Lasso regularization, the rest blood pressure and the abnormality of ST-T segment in ECG analysis do not play an important part. By contrast, in the model using the Ridge approach, these did contribute in the model.

Table IV depicts accuracy performance of two models during training period and hold-out test stage. We found that the performance of both models was consistent and greater than 70% through training and test sets.

IV. Discussion

In this work, we have examined the possibility of reducing the number of risk factors, especially those require costly and invasive laboratory-based results, to detect heart disease. We found that age, gender, rest blood pressure measurement, maximum heart rate, and the abnormality of ST segment in the rest ECG can be used to feed into a simple generalized linear model and achieved closely comparable accuracy as earlier works that utilized a more comprehensive input set. For example, authors of the well-known work [4] that included laboratory based data only yielded approximately a 77% classification accuracy (using a logistic regression approach). Furthermore, research advances in processing massive datasets may provide a useful real-time tool and massive information learning platform that cardiologists can assess an individual patient's risk for heart disease more accurately with less laboratory cost and faster. It is worth noting that the above ECG data used in our method can be obtained easily given the recent advances in wearable sensor for automated ECG analysis.

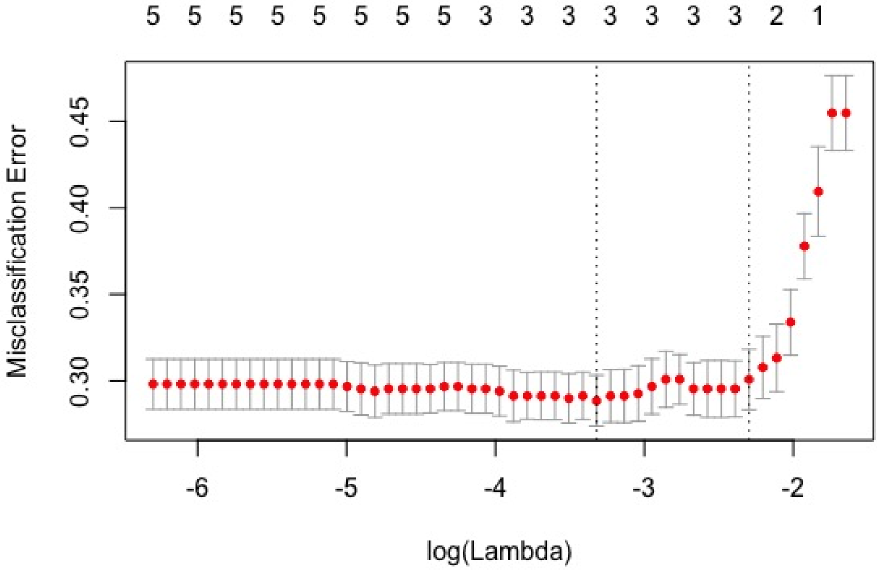
Relevancy of each component of the proposed information set has been long supported in clinical studies. Aging has been suggested one of the highest risk factors for CAD



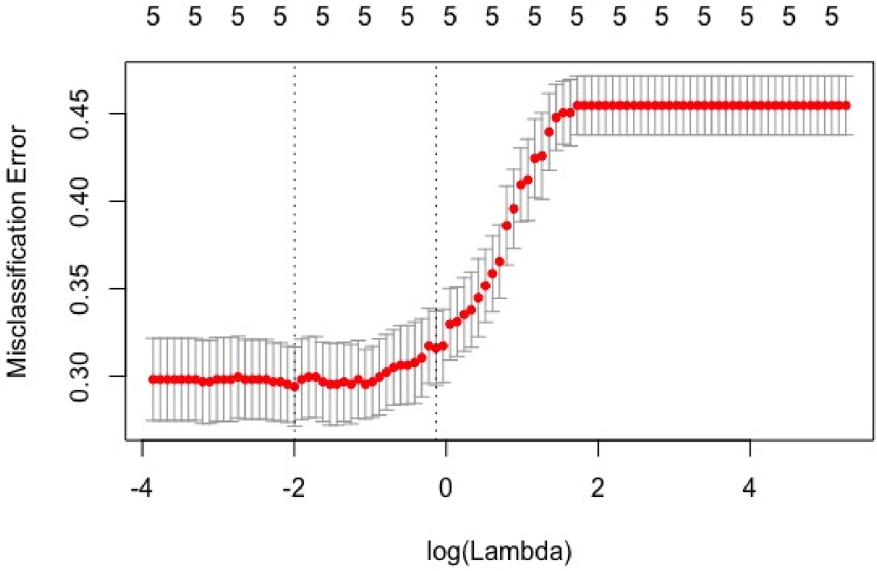
(a) Fitted GLM with Lasso penalty.

(b) Fitted GLM with ridge penalty.

Fig. 1. Fitted generalized linear models with regularization paths.



(a) Cross-validation for GLM with Lasso penalty.



(b) Cross-validation for GLM with ridge penalty.

Fig. 2. Cross-validation for generalized linear models (GLM) with different regularization paths. The top axis denotes the number of non-zero coefficients at a given λ. λ_{min} (λ_{SE}) are depicted by dotted lines. Error bars are upper and lower standard deviations.

TABLE I

MULTI-CLASS DISTRIBUTION (N PEOPLE) AND PREVALANCE % OF *Absence =Negative* OR *Presence= Positive* AMONG DATABASE LOCATIONS.

Database	Multi-Class (n people)					Total	Class Prevelance (%)	
	0	1	2	3	4		Negative	Positive
Cleveland	164	55	36	35	13	303	54.1%	45.9%
Hungarian	188	37	26	28	15	294	63.9%	36.1%
Switzerland	8	48	32	30	5	123	6.5%	93.5%
California	51	56	41	42	10	200	25.5%	74.5%

TABLE II

BASELINE CHARACTERISTICS OF 13 RISK FACTORS (THE WHOLE DATASET).

Factors		Class		P-value
		Negative	Positive	
		(n=411)	(n=509)	
Age		50.547 (9.4)	55.904 (8.7)	< 0.001
Sex	Men	267 (65 %)	459 (90.2 %)	< 0.001
	Women	144 (35 %)	50 (9.8 %)	
Chestpain (yes)	Typical	26 (6.3 %)	20 (3.9 %)	< 0.001
	Atypical	150 (36.5 %)	24 (4.7 %)	
	Non-aginal	131 (31.9 %)	73 (14.3 %)	
	Asymptomatic	104 (25.3 %)	392 (77.0 %)	
Rests blood pressure		129.9 (16.87)	133.979 (20.552)	0.002
Cholesteral		227.9 (75.8)	176.48(127.518)	< 0.001
Fasting blood sugar $> 120mg/dl$		44 (11.1 %)	94 (21.7 %)	< 0.001
RestECG	Normal	268 (0.652)	283 (0.558)	0.003
	Abnormal ST	61 (0.148)	118 (0.233)	
	LVHypertrophy	82 (0.2)	106 (0.209)	
Max heart rate		148.8 (23.6)	128.262 (24.02)	< 0.001
CPETAgina (yes)		55 (14.1 %)	282 (59.55 %)	< 0.001
Oldpeak		0.418 (0.716)	1.263 (1.197)	< 0.001
Slope		1.49 (0.62)	1.93 (0.56)	< 0.001
CA		0.279 (0.640)	1.132 (1.012)	< 0.001
Thal		3.99 (1.68)	5.92 (1.656)	< 0.001

TABLE III

COEFFICIENTS OF FITTED MODELS (AT THE BEST CROSS-VALIDATION PARAMETER λ_{min}) FOR FIVE NON-LABORATORY-BASED RISK FACTORS.

Risk factors	Coef. by Lasso [a]	Coef. by Ridge [b]
Age	0.029	0.029
RestBP	0	0.003
MaxHR	-0.021	-0.016
SexMale	0.998	0.871
Rest ECG normal-ST	0	-0.187

[a] Coefficients estimated by GLM with Lasso regularization
[b] Coefficients estimated by GLM with Ridge regularization

[5], [6], [9]. Authors of [9] showed the genetic relationship between aging and heart disease. Comparing with commonly used laboratory-based risk scores: Atherosclerotic Cardiovascular Disease (ASCVD) [10], Framingham Risk Score (FRS) [11], and SCORE (Systematic Coronary Risk Evaluation) [12] a non-laboratory-based risk tool [13] has shown to have a very high correlation (N=47,466 people, cross-sectional collection from nine countries). More recently, non-laboratory Framingham score [14], which substitute BMI for lipids in FRS [11], was shown as the best performance among non-laboratory algorithms (internal validity only). However, these non-laboratory scores only eliminated blood-test based factors while maintain a larger number of inputs than our proposed model. Hence, using our approach, *Big data* based systems can be utilized for heart disease detection without laboratory-based values. This risk assessment approach is applicable to population where laboratory testing is not easily accessible (e.g., developing countries or regions with limited resources).

REFERENCES

[1] World Health Organization, “Cardiovascular diseases (cvds),” May 2017.

[2] Thomas A Gaziano, Cynthia R Young, Garrett Fitzmaurice, Sidney Atwood, and J Michael Gaziano, “Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk:

TABLE IV

CONFUSION MATRIX AND STATISTICS WHEN USING FIVE NON-LABORATORY-BASED RISK FACTORS DURING TRAINING PERIOD AND HOLD-OUT TEST STAGE.

	Training Performance		Hold-out Test	
	Lasso	Ridge	Lasso	Ridge
Accuracy	0.7170	0.7060	0.7252	0.7405
95% CI	(0.6828, 0.7495)	(0.6715, 0.7389)	(0.6404, 0.7995)	(0.6566, 0.8131)
No Information Rate	0.5453	0.5453	0.5420	0.5420
P-Value [Acc > NIR]	< 0.0001	< 0.0001	< 0.0001	< 0.0001
Kappa	0.4227	0.4003	0.4392	0.4690
Mcnemar's Test P-Value	0.0017	0.0021	0.1336	0.0592
Sensitivity	0.7985	0.7884	0.8169	0.8451
Specificity	0.6193	0.6073	0.6167	0.6167
Pos Pred Value	0.7156	0.7065	0.7160	0.7229
Neg Pred Value	0.7193	0.7053	0.7400	0.7708
Prevalence	0.5453	0.5453	0.5420	0.5420
Detection Rate	0.4354	0.4299	0.4427	0.4580
Detection Prevalence	0.6085	0.6085	0.6183	0.6336
Balanced Accuracy	0.7089	0.6978	0.7168	0.7309

the NHANES I follow-up study cohort," *The Lancet*, vol. 371, no. 9616, pp. 923 – 931, 2008.

[3] Jisheng Cui, Andrew Forbes, Adrienne Kirby, John Simes, and Andrew Tonkin, "Laboratory and non-laboratory-based risk prediction models for secondary prevention of cardiovascular disease: the lipid study," *European Journal of Cardiovascular Prevention & Rehabilitation*, vol. 16, no. 6, pp. 660–668, 2009.

[4] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H. Guppy, Stella Lee, and Victor Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *The American Journal of Cardiology*, vol. 64, no. 5, pp. 304 – 310, 1989.

[5] George A Diamond and James S Forrester, "Analysis of probability as an aid in the clinical diagnosis of coronary-artery disease," *New England Journal of Medicine*, vol. 300, no. 24, pp. 1350–1358, 1979.

[6] Frederick K Korley, Constantine Gatsonis, Bradley S Snyder, Richard T George, Thura Abd, Stefan L Zimmerman, Harold I Litt, and Judd E Hollander, "Clinical risk factors alone are inadequate for predicting significant coronary artery disease," *Journal of Cardiovascular Computed Tomography*, 2017.

[7] Harry Hemingway, Gene S Feder, Natalie K Fitzpatrick, Spiros Denaxas, Anoop D Shah, and Adam D Timmis, "Using nationwide big datafrom linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the clinical disease research using linked bespoke studies and electronic health records (caliber) programme," 2017.

[8] Jerome Friedman, Trevor Hastie, and Rob Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software, Articles*, vol. 33, no. 1, pp. 1–22, 2010.

[9] Brian J. North and David A. Sinclair, "The intersection between aging and cardiovascular disease," *Circulation Research*, vol. 110, no. 8, pp. 1097–1108, 2012.

[10] Paul Muntner, Lisandro D Colantonio, Mary Cushman, David C Goff, George Howard, Virginia J Howard, Brett Kissela, Emily B Levitan, Donald M Lloyd-Jones, and Monika M Safford, "Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations," *Jama*, vol. 311, no. 14, pp. 1406–1415, 2014.

[11] Keaven M Anderson, PW Wilson, Patricia M Odell, and William B Kannel, "An updated coronary risk profile. a statement for health professionals.," *Circulation*, vol. 83, no. 1, pp. 356–362, 1991.

[12] R.M. Conroy, , K. Pyrl, , A.P. Fitzgerald, , S. Sans, , A. Menotti, , G. De Backer, , D. De Bacquer, , P. Ducimetire, , P. Jousilahti, , U. Keil, , I. Njlstad, , R.G. Oganov, , T. Thomsen, , H. Tunstall-Pedoe, , A. Tverdal, , H. Wedel, , P. Whincup, , L. Wilhelmsen, , I.M. Graham, and , "Estimation of ten-year risk of fatal cardiovascular disease in europe: the score project," *European Heart Journal*, vol. 24, no. 11, pp. 987–1003, 2003.

[13] Thomas A. Gaziano, Shafika Abrahams-Gessel, Sartaj Alam, Dewan Alam, Mohammed Ali, Gerald Bloomfield, Rodrigo M. Carrillo-Larco, Dorairaj Prabhakaran, Laura Gutierrez, Vilma Irazola, Naomi S. Levitt, J. Jaime Miranda, Antonio Bernabe-Ortiz, Ankur Pandya, Adolfo Rubinstein, Krisela Steyn, Denis Xavier, and Lijing L. Yan, "Comparison of nonblood-based and blood-based total cv risk scores in global populations," *Global Heart*, vol. 11, no. 1, pp. 37 – 46.e2, 2016, Investment in Global Health Research: A Public Private Partnership.

[14] Jacob K Kariuki, Eileen M Stuart-Shor, Suzanne G Leveille, Philimon Gona, Jerry Cromwell, and Laura L Hayman, "Validation of the nonlaboratory-based framingham cardiovascular disease risk assessment algorithm in the atherosclerosis risk in communities dataset," *Journal of Cardiovascular Medicine*, vol. 18, no. 12, pp. 936–945, 2017.